The test of Lasègue: systematic review of the accuracy in diagnosing herniated discs.

Devillé, W.L.J.M.; van der Windt-Mens, D.A.W.M.; Dzaferagic, A.; Bezemer, P.D.; Bouter, L.M.

published in
Spine
2000

DOI (link to publisher)
10.1097/00007632-200005010-00016

document version
Publisher's PDF, also known as Version of record

SPINE Volume 25, Number 9, pp 1140–1147

The Test of Lasègue

Systematic Review of the Accuracy in Diagnosing Herniated Discs

Walter L. J. M. Devillé, MD, DTMH, MSc,* Daniëlle A. W. M. van der Windt, PhD,* Aida Džaferagić, MSc,† P. D. Bezemer, PhD,* and Lex M. Bouter, PhD*

Study Design. A systematic review of the literature including statistical meta-analysis.

Objectives. To evaluate published methods of the test of Lasègue or straight leg raising test and the cross straight leg raising test by using a recently developed criteria list and to summarize and explore reasons for variation in diagnostic accuracy.

Summary of Background Data. Little evidence exists on the diagnostic accuracy of the widely used straight leg raising test and the cross straight leg raising test in diagnosing herniated discs in patients with low back pain.

Methods. MEDLINE and EMBASE searches up to 1997 showed 17 diagnostic publications evaluating the straight leg raising test with surgery as reference standard. Quality of methods was assessed with a specific checklist. Eleven studies were selected for statistical pooling. Sources of variation and heterogeneity were studied by meta-regression of the diagnostic odds ratio.

Results. All studies were surgical case-series at nonprimary care level. Verification-bias was obvious in one study. Pooled sensitivity for straight leg raising test was 0.91 (95% CI 0.82–0.94), pooled specificity 0.26 (95% CI 0.16–0.38). Pooled diagnostic odds ratio was 3.74 (95% CI 1.2–11.4). Discriminative power was lower in recent studies, in studies with only inclusion of primary hernias, and with blind assessment of both the index-test (straight leg raising test) and the reference (surgery). For the cross straight leg raising test pooled sensitivity was 0.29 (95% CI 0.24–0.34), pooled specificity was 0.88 (95% CI 0.86–0.90), and the pooled diagnostic odds ratio 4.39 (95% CI 0.74–25.9).

Conclusions. The diagnostic accuracy of the straight leg raising test is limited by its low specificity. Discriminative power decreased with a more valid design, a more homogenous case-mix, and year of publication. Although the studies may reflect everyday clinical practice, they do not enable a valid evaluation of the diagnostic accuracy of both tests. Diagnostic research should evaluate the validity of the complete diagnostic process and study the evidence of the added value of the different tests used. [Key words: sensitivity, specificity, diagnosis, meta-analysis, test of Lasègue, straight leg raising test] **Spine 2000;25: 1140–1147**

Although approximately 70% of the adult population experiences low back pain once or more during their life, no specific pathology is identified in up to 85% of the patients.[9] Approximately 1.5% of low back pain patients endure symptoms of sciatica, and only approximately 2% undergo surgery.[9,10] Associations of herniated discs with signs and symptoms and even imaging results remain weak.[45,33] It must be realized that herniated discs can be found by imaging diagnostic tests in 20% to 30% of symptom-free persons.[5,44]

Once systemic diseases are excluded as possible causes of low back pain, careful diagnostic neurologic evaluation remains important to avoid unnecessary surgical interventions.[11] The straight leg raising test (SLR), also known as the test of Lasègue, and the cross straight leg raising test (CSLR), are two tests based on stretching of the nerves in the spine. The SLR frequently is used in primary care for making decisions about diagnostic imaging or hospital referral.[19]

A few recent reviews have discussed the value of history taking and physical examination for diagnosing herniated discs in patients with low back pain, but only two reviewed the available literature in a systematic manner.[3,11,18,41] Both of the latter two reviews were criteria-based, but only one offered a quantitative summary of the findings.[41] No review tried to study the variability of the diagnostic accuracy of the (C)SLR. The most recent review concluded that the test of Lasègue had a high sensitivity and a low specificity, but that it varied greatly across studies.[41]

The present review was conducted as an update for the (C)SLR up to 1997 and to carefully reassess methodologic quality by using a recently developed criteria list.[18,22] It contains a more detailed analysis by summarizing the diagnostic accuracy quantitatively and exploring reasons for its variability in the diagnosis of herniated discs.[40] The methods of this diagnostic meta-analysis closely adhere to recently developed guidelines for conducting diagnostic reviews.[20,21,31,34]

Methods

Literature Search. The literature search of an earlier review was extended from 1992 through 1997 using MEDLINE and EMBASE databases (key words and free-text words screening title, subtitle, abstract, and key words [radiculopathy, backache, low back, Lasègue, straight leg raising, and cross straight leg raising]) and also screened the references from identified publications.[18] The inclusion and exclusion criteria from the earlier review were used: studies using surgery as reference standard, studies presenting at least data on the sensitivity only or on both sensitivity and specificity of one or both tests were included and review articles and studies including fewer than 10 patients with disease were excluded. In addition, two pub-

From the *Department of Clinical Epidemiology and Biostatistics and Research Institute for Research in Extramural Medicine (EMGO Institute), Vrije Universiteit Amsterdam, and the †Universiteit Utrecht, The Netherlands.
Acknowledgment date: December 15, 1998.
First revision date: April 28, 1999.
Acceptance date: August 6, 1999.
Device status category: 1.
Conflict of interest category: 12.

Table 1. Criteria of Internal and External Validity for Diagnostic Studies

Criteria of Internal Validity (IV)		Positive Score
1	Valid reference standard	Surgery
2	Definition of cut-off point for reference standard	Type of hernia given (protruded, extruded, and/or sequestrated)
3	Blind measurement of index test and reference test	In both directions or only index or reference test
4	Avoidance of verification bias	Assessment by reference standard independent from index test results
5	Index test interpreted independently of all clinical information	Mentioned in the publication
Criteria of External Validity (EV)		**Positive Score**
1	Spectrum of disease	In- and/or exclusion criteria mentioned
2	Spectrum of nondisease	Details given
3	Setting	Enough information to identify setting (community through tertiary care)
4	Duration of illness before diagnosis	Mentioned
5	Previous tests/referral filter	Details given about clinical and other diagnostic information as to which the index test is being evaluated
6	Comorbid conditions	Details given in diseased and nondiseased
7	Demographic information	Age and/or gender details given
8	Execution of index-test	Information about standard procedure directly or indirectly available
9	Explicitation of the cut-off point of index-test	Degree of leg elevation
10	Percentage excluded	If appropriate: exclusions mentioned (infeasibility or indeterminate results)
11	Reproducibility	Studied

lications on Cauda Equina syndrome from the first set of publications were excluded, as this emergency syndrome, which needs urgent surgery, has its own pathology diagnosed by a specific history and specific symptoms. One researcher (W.D.) selected the studies.

Quality and Applicability of Studies. The checklist of the Cochrane Methods Working Group on Meta-analysis of Diagnostic and Screening Tests was used to assess the quality of methods of selected studies.[22] Two items from the list of Mulrow[32] were added to the checklist: cutoff point of the reference test and purpose of the test (screening, case-finding or diagnosis). Two reviewers (W.D., A.D.) independently assessed all selected publications. Disagreements were solved in a consensus meeting. Agreement between both reviewers was quantified by Cohen's *kappa* (κ).[14]

Detailed guidelines for the assessment of each item were made available to the reviewers (Table 1). Internal validity criteria (IV) were scored as "positive" (adequate methods), "negative" (inadequate methods, potential bias), or unclear if insufficient information had been provided on a specific item. External validity criteria (EV) were assessed to evaluate generalizibility. These criteria scored positive if sufficient information was provided to judge generalizibility of findings. After the consensus meeting we decided to score unclear scores as negative. Subtotals were calculated for internal (maximum 6) and external validity (maximum 11) separately.

When information about blinding of measurements (IV2) was not given in surgical case series, blind assessment of the index test *versus* the reference test was assumed, but not *vice versa*. If the information could not be derived from the publication, verification bias (IV3) was classified as possible in cases of retrospective cohorts of surgical case series. If clinical information (IV4) remained unclear from the publication, this point was assessed as negative if the study was based on daily clinical practice.

Spectrum of disease (EV1) was defined as small if only upper (L1–L3) or lower disc hernias were included. Spectrum of nondisease (EV2) was defined as broad if the group of patients without disc hernia was constituted by more than one clinical diagnosis. Explicitation of the cutoff point (EV9): a typical positive passive SLR reproduces the sciatic pain between 30° and 60–75°.[3,11,43] An atypical SLR was defined by three studies as pain produced in the back only.[12,27,28]

In addition to the abovementioned criteria, information was collected on year of study, disease prevalence at the setting, sample size and size of the smallest group (with disease or without disease), previous surgery or bilateral sciatica, level of hernia, detailed information on the group without disease, and if the study population was collected prospectively or retrospectively. Where possible, information was collected on a numerical scale.

Meta-analysis. Data on sensitivity and specificity of SLR and CSLR for detecting disc hernias were derived from the original numbers given in the publications to avoid rounding-off effects. If absolute numbers were not presented, sensitivity and specificity were used as presented. In one study, the required data were derived from a receiver operating characteristic (ROC) curve, plotting sensitivity against (1 − specificity) at different cutoff points.[1] Only studies presenting data on both parameters were selected for statistical pooling of results. Sensitivity and specificity were pooled after natural logarithmic transformation.[3] The average predictive values were calculated based on the pooled estimates of sensitivity and specificity at the mean prevalence of the pooled studies.[3] The diagnostic odds ratio (DOR) of each individual study—a measure for the discriminative power of the test—was calculated according to the following formula[31,34]:

$$\mathrm{DOR} = \frac{\text{sensitivity}/(1 - \text{sensitivity})}{(1 - \text{specificity})/\text{specificity}}$$

The DOR calculates the ratio of the odds of a positive test result in the patient with disease on the odds of a positive test result in the patient without disease. A DOR of 1 means that the odds of a positive test result in the study group with disease and the study group without disease is equal, and the test has no discriminative power. When the DOR is more than one, the odds of a positive test result is higher in the population with disease.

Table 2. Diagnostic Studies on the (Cross) Straight-Leg-Raising Test Included in the Systematic Review

Studies	Year	Prior Probability	Age (yr)	M/F Ratio	Lower/Upper Lumbar Hernia Ratio	Duration (weeks)	Exclusion	
							Previous surgery	Bilateral
Charnley	1951	—	—	—	—	—	No	No
Gurdjian	1961	0.98	41	1.9	51.3	—	No	No
Knuttson	1961	0.90	42	1.5	60.3	44	Yes	No
Aronson	1963	—	52	—	0	—	No	No
Hirsch	1963	0.77	42	1.8	—	—	Yes	No
Hakelius	1972	0.75	—	—	—	—	No	No
Spangfort	1972	0.86	41	2.4	46.9	170	No	No
Edgar	1974	—	38	1.5	—	21	No	No
Kosteljanetz	1984	0.58	—	1.0	—	16	Yes	Yes
Kortelainen	1985	—	41	1.6	39.3	12	Yes	No
Shiqing	1987	—	35	6.1	—	—	No	No
Kerr	1988	0.74	40	1.2	32.3	21	No	No
Kosteljanetz	1988	0.87	45	1.5	—	—	Yes	Yes
Albeck	1996	0.76	40	1.5	—	—	Yes	No
Jönsson	1996	0.91	43	1.5	20.1	13	Yes	No

Age: mean or median; exclusion: previous hernia surgery and/or bilateral radiculopathy.

Pooling of the DOR also was performed after natural logarithmic transformation [ln(DOR)].

The statistical heterogeneity of sensitivity and specificity across studies was tested by a chi-square test of independency with $\kappa - 1°$ of freedom (κ = number of studies).[21,29] Heterogeneity of the DOR was tested using the test mentioned by DerSimonian and Laird.[8]. In case of no statistically significant heterogeneity, pooling was carried out according to the fixed effect model.

Using meta-regression, the influence on diagnostic accuracy of different sources of variation was evaluated. The cutoff point of the index-test used in the different studies, all validity criteria and the validity scores, and numerical information about study characteristics and study population were used as independent variables in univariate meta-regression analysis. Multivariate meta-regression was not performed because of the limited number of studies. Both unweighted and weighted regression was conducted using the inverse of the variance of the DOR as a weighing factor. As the validity of this weighing is still doubtful in the meta-analysis of diagnostic studies,[13] only the results of the unweighted pooling are presented. A regression line was fitted as a Summary ROC curve (SROC) in a scatter plot of the different included studies with their sensitivity on the y-axis and (1 − specificity) on the x-axis.[31,34] Estimates are presented with 95% confidence intervals (95% CI). Analyses were performed with SPSS 7.5 for Windows95 (Chicago, IL) and with Meta-test.[30]

■ Results

Literature Search

The systematic review published in 1995 retained 19 studies on radiculopathy from 540 retrieved publications. Twelve publications reporting diagnostic parameters about the (C)SLR were selected for this review. The additional search up to 1998 revealed 12 more studies, of which 3 were added to the first selection. Among the studies excluded, three presented only data on the reproducibility of the SLR, on predictive value, or combined in models. Five publications were reviews.[2,3,11,18,41]

Quality and Applicability of Studies

Table 2 presents the 15 selected studies[4,6,12,15–17,23–28,37,38,42] and some characteristics of the study population. Median age of the study population was 40.9 years (range, 35–52 years), median male/female ratio was 1.5 (range, 1–6), and median duration before diagnosis was 21 weeks (range, 12–170). One study was limited to upper lumbar L1–L3 pathology, five to lower L4–S1 pathology, and six studies included the whole lumbosacral region. Four studies evaluated both typical and atypical SLRs. Eight studies provided data about the CSLR.

Table 3 presents the results of the quality assessment. All patients with disc hernias had surgery as a reference standard (IV1). Definitive hernia was defined as extruded, protruded and bulging disc or sequestrated in 11 studies, as extruded and protruded in 2 other studies, and as extruded only in 1 study (IV2). Information was insufficient to assess verification bias in the other 14 studies. Two studies had a blind assessment in both directions (IV3). In one study, the group without disease did not have surgery, but was selected on the basis of a normal myelography, and verification bias was obvious (IV4).[24] Six studies were prospective.

Regarding the external validity, 11 studies included patients without a hernia in the study, but only 5 gave details about these controls (EV2). No study was done in a primary care setting (EV3). Five studies gave details about the procedure used and four mentioned the standard supine procedure with a hand on the iliac rim (n = 1) or the knee (n = 2; EV8). Six studies mentioned the cutoff point used: $<70°$ (n = 3), $<80°$ (n = 1), or $<90°$ (n = 2) (EV9). The added value of neck or foot dorsiflexion was mentioned four times, but not evaluated. Two studies mentioned tight hamstrings without details. One study gave details about patients excluded for the test evaluation (EV10). The other studies seemed to include

Table 3. Quality Assessment of Diagnostic Studies Included in the Review

Author	Year	Sample	2by2	IV1	IV2	IV3	IV4	IV5	IV Total	EV Total	% of Maximum Validity IV + EV
Edgar	1974	50	−	+	+	+	?	+	4	7	65%
Kosteljanetz	1988	55	+	+	+	+	?	−	3	8	65%
Kosteljanetz	1984	100	+	+	+	+	?	?	3	7	59%
Gurdjian	1961	1176	+	+	+	+	?	+	4	6	59%
Jönsson	1995	150	+	+	+	+	?	?	3	7	59%
Shiqing	1987	113	−	+	+	+	?	?	3	6	53%
Spangfort	1972	2504	+	+	+	+	?	?	3	5	47%
Kerr	1988	136	+	+	+	+	−	−	3	5	47%
Knuttson	1961	182	+	+	−	+	?	−	2	5	41%
Aronson	1963	73	−	+	+	+	?	?	3	4	41%
Kortelainen	1985	470	−	+	+	++	?	−	4	3	41%
Albeck	1996	80	+	+	+	++	?	−	4	2	35%
Charnley	1951	88	+	+	−	+	?	?	2	3	29%
Hakelius	1972	1986	+	+	+	+	?	−	3	2	29%
Hirsch	1963	232	+	+	−	+	?	−	2	3	29%

IV = criterium of internal validity; EV = external validity (see table 1); ? = information unclear; 2by2 = diagnostic table available.
+ Positive score; − no score (see Table 1).

all patients. Eight studies looked at the CSLR, four of which studied a smaller number of patients.

The median score for internal validity was 3/6 (50%; range, 33–66%) and 5/11 for external validity (45%; range, 18–72%). Only four studies had a score of 4 in 6 for internal validity. Median total validity was 47% (range, 29–65%). Six studies of 15 scored 50% or more for the total validity.

Median overall agreement between both assessors was 76.5% (range, 32.9–100%) with a κ of 0.56 (range, 0.33–1) for all criteria. Agreement was relatively poor for verification bias ($\kappa = 0.33$), inclusion/exclusion criteria ($\kappa = 0.34$), spectrum of "nondiseased" ($\kappa = 0.35$), and exclusion of test results ($\kappa = 0.37$). All disagreements were resolved during the consensus meeting. Disagreements were because of reading errors, differences in interpretation, and experience.

Meta-analysis and Meta-regression

The results of 11 studies reporting a complete diagnostic 2-×-2 table were available for statistical pooling (Table 4). Six studies provided sufficient data to enable a meta-analysis of the CSLR.

Straight Leg Raising Test. Tests of heterogeneity were highly significant for both sensitivity and specificity. weighted random effect (REM) pooled sensitivity was 0.91 (95% CI 0.82–0.94), and pooled specificity was 0.26 (95% CI 0.16–0.38).

The DOR across all studies was heterogenous ($Q = 20.65$; 10 *df*; $0.025 > P > 0.01$). The unweighted pooled DOR was 3.74 (95% CI 1.23–11.4). By excluding the study of Kerr et al,[24] which was a clear outlier with a DOR of 39.2 (Figure 1), the DOR across studies became homogeneous ($Q = 8.8$; 9 *df*; $P > 0.25$). The pooled DOR after the exclusion of the outlier was 2.96 (95% CI 2.05–4.28). This outlier was the only study that used a nonsurgical group of patients as the cases without disease. It was excluded from further analysis of covariates and subgroups in the meta-regression analysis. Figure 1 shows the DOR on a logarithmic scale with 95% confidence intervals for the individual studies, as well as the pooled DOR including and excluding the outlier study. The mean predictive value of a positive result at the average prevalence of 0.86 in these studies was 0.89. The negative predictive value was 0.33.

The meta-regression revealed no association between the ln(DOR) and the cutoff point of a positive SLR used in the different studies, but showed a strong negative

Table 4. Parameters of Diagnostic Accuracy of the (Cross) Straight-Leg-Raising Test for the Detection of Disc Hernia (CSLR and SLR)

	Diseased +	Diseased −	Sensitivity (95%CI)	Specificity (95%CI)	DOR (95%CI)
SLR					
Edgar	50	−	0.80 (0.79–0.91)	—	—
Kosteljanetz	45	7	0.89 (0.75–0.96)	0.14 (0.01–0.58)	1.33 (0.20–8.83)
Kosteljanetz	58	42	0.78 (0.64–0.87)	0.48 (0.32–0.63)	3.15 (1.35–7.36)
Gurdjian	1151	25	0.81 (0.78–0.83)	0.52 (0.32–0.72)	4.53 (2.07–9.93)
Jönsson	178	18	0.87 (0.81–0.91)	0.22 (0.07–0.48)	1.86 (0.60–5.75)
Shiqing	110	−	0.94 (0.90–0.98)	—	—
Spangfort	2157	347	0.97 (0.96–0.97)	0.11 (0.08–0.15)	3.83 (2.55–5.76)
Kerr	100	36	0.98 (0.92–1.00)	0.44 (0.28–0.62)	39.20 (10.4–147)
Knuttson	186	20	0.95 (0.91–0.98)	0.10 (0.02–0.33)	2.19 (0.54–8.88)
Aronson	73	−	0.40 (0.29–0.51)	—	—
Kortelainen	403	−	0.94 (0.92–0.96)	—	—
Albeck	61	19	0.82 (0.70–0.90)	0.21 (0.07–0.46)	1.21 (0.36–4.06)
Charnley	74	14	0.85 (0.75–0.92)	0.57 (0.30–0.81)	7.64 (2.31–25.3)
Hakelius	1467	492	0.96 (0.95–0.97)	0.14 (0.11–0.18)	4.18 (2.90–6.02)
Hirsch	179	53	0.91 (0.85–0.94)	0.32 (0.20–0.46)	4.50 (2.17–9.52)
Pooled estimate:			0.91 (0.82–0.94)	0.26 (0.16–0.38)	3.97 (3.22–4.9)
CSLR					
Edgar	50	−	0.44 (0.79–0.91)	—	—
Kosteljanetz	19	1	0.57 (0.34–0.79)	1.00 (0.03–1.00)	4.06 (0.12–135)
Jönsson	150	14	0.22 (0.16–0.30)	0.93 (0.64–1.00)	3.67 (0.75–17.8)
Shiqing	110	−	0.15 (0.90–0.98)	—	—
Spangfort	2157	342	0.23 (0.21–0.25)	0.88 (0.84–0.91)	2.21 (1.58–3.11)
Kerr	100	36	0.43 (0.33–0.53)	0.93 (0.80–0.99)	12.82 (3.64–45.1)
Knuttson	162	20	0.25 (0.18–0.32)	0.93 (0.73–1.00)	6.22 (1.35–28.8)
Hakelius	1467	492	0.28 (0.25–0.30)	0.88 (0.84–0.90)	2.71 (2.03–3.63)
Pooled estimate:			0.29 (0.24–0.34)	0.88 (0.86–0.90)	4.39 (0.74–25.9)

DOR = diagnostic odds ratio.

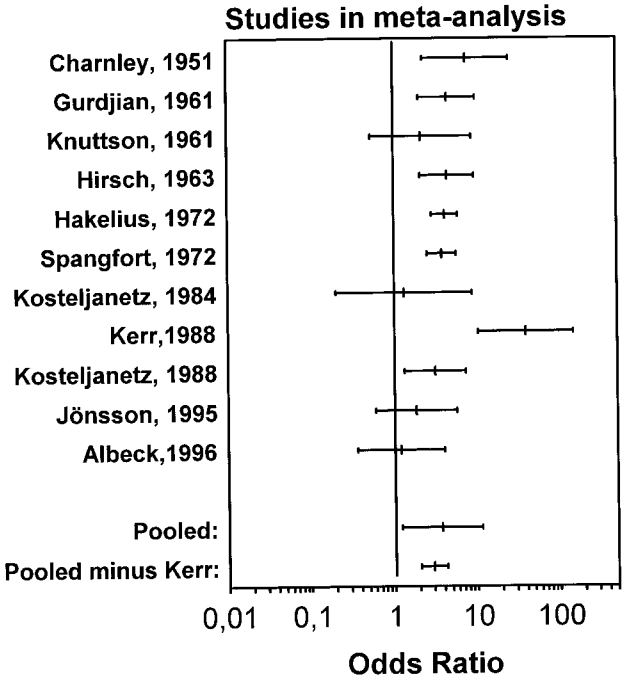

Figure 1. Diagnostic odds ratio (DOR) of the straight leg raising test (SLR) with 95% confidence intervals for studies included in the statistical meta-analysis, ordered by year of publication. Pooled DOR with and without outlier study.

correlation with year of publication. Small differences were found for studies excluding patients who previously had had surgery and when index-test and reference test were assessed independently from each other (Table 5). The DOR of the SLR was lower in recently published studies, in studies with only primary hernias, and with blind assessment of both the index-test (SLR) and the reference (surgery).

The verification bias in the excluded study, suspected during the methods assessment, was confirmed by meta-regression (Figure 1; Table 5)[24]; the DOR or discriminative power was higher compared with studies without apparent verification bias.

Table 5. Selection of Covariates in Univariate Metaregressions of the Diagnostic Odds Ratio (DOR) of the Straight-Leg-Raising Test (SLR)

	Regression Coefficient (95% CI)	*P*
Covariate		
Cut-off point of SLR	−0.0958 (−0.389–0.198)	0.909
Blind assessment	−0.8860 (−2.043–0.271)	0.113
Verification bias	2.6890 (1.418–3.959)	0.001
Year of publication	−0.0308 (−0.048–−0.013)	0.004
Exclusion of previously operated patients	−0.6730 (−1.388–0.043)	0.062

		Pooled DOR (95%CI)
Analysis		
Unweighted	Outlier in	3.74 (1.23–11.4)
	Outlier out	2.96 (2.05–4.28)
Weighted	Outlier in	3.97 (3.22–4.90)
	Outlier out	3.74 (3.00–4.66)

Pooled estimates of the DOR.

Figure 2 shows the most important differences between subgroups. There was no association with any of the other covariates or study characteristics nor with the total number of validity items scored positive. The pooled DOR of the highest tertile of total validity was not statistically different from the lowest tertile. Studies in the highest tertile for external or internal validity showed a lower DOR compared with the lowest tertile, but differences were not statistically significant.

Figure 3 shows the fitted curve from the meta-regression in a summary ROC area. The area under the fitted curve (AUC) is 0.70 if the fitted curve is extrapolated to the axes (thin line on the figure).

Cross Straight Leg Raising Test. Tests of heterogeneity were significant for both sensitivity and specificity. The weighted random effect (REM) pooled sensitivity was 0.29 (95% CI 0.24–0.34), and the pooled specificity was 0.88 (95% CI 0.86–0.90).

The DOR across all studies was homogeneous ($Q = 8.5$; $5\ df$; $0.25 > P > 0.10$). The unweighted pooled DOR was 4.39 (95% CI 0.74–25.9; Table 4). The DOR was not associated with the cutoff point, but was different between the study with the verification bias (DOR 12.8) and the other five studies (DOR 3.32, 95% CI 2.35–4.69; $P = 0.035$).

The predictive value of a positive test was 0.92 at a prevalence of 0.82 in these six studies. The negative predictive value was 0.22.

■ Discussion

The diagnostic accuracy of the SLR in detecting disc hernia seems to be limited by its low specificity. Although the sensitivity is high, the diagnostic odds ratio remains low as the probability of a positive SLR in surgical patients without a disc hernia remains high. This is probably because of the casemix in these referred and selected patient populations. These patients, who have all received surgery, are at the severe end of the pathologic spectrum. This results also in a high prevalence and, consequently, in a high predictive value of a positive SLR.

It is preferable not to draw final conclusions about the diagnostic accuracy based on biased studies. Therefore, a thorough assessment of the validity of all selected studies was included. The criteria-list used by the authors of the previously published systematic review was not used because it was based on a scoring system with weighted items.[18] There is considerable debate about the use of arbitrarily chosen weights for different items in these lists, especially in observational studies.[7,39] The criteria-list used in the present study was developed by the Cochrane Diagnostic and Screening Tests Working Group and explicitly makes a clear distinction between internal validity and external validity items.[22] This criteria list includes verification bias as a separate item and evaluates

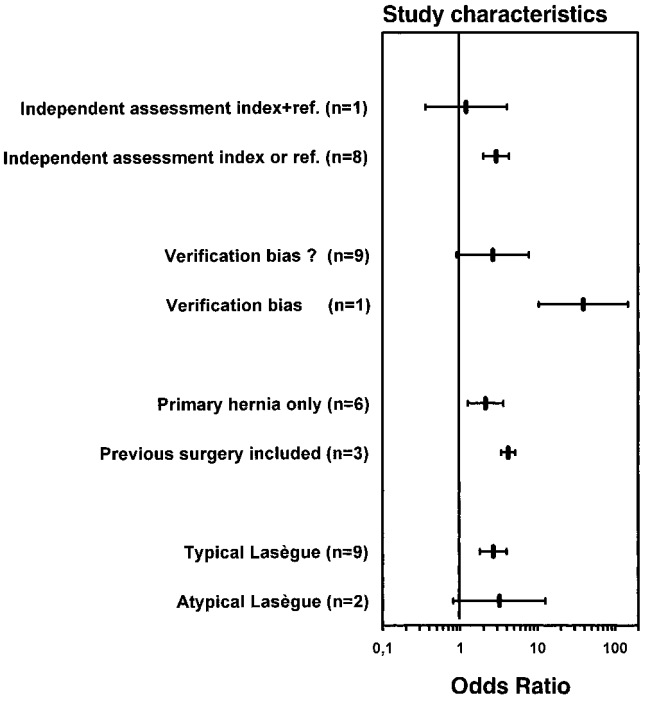

Figure 2. Different study and population characteristics used in subgroup-analysis and their respective DOR of the SLR with 95% confidence intervals.

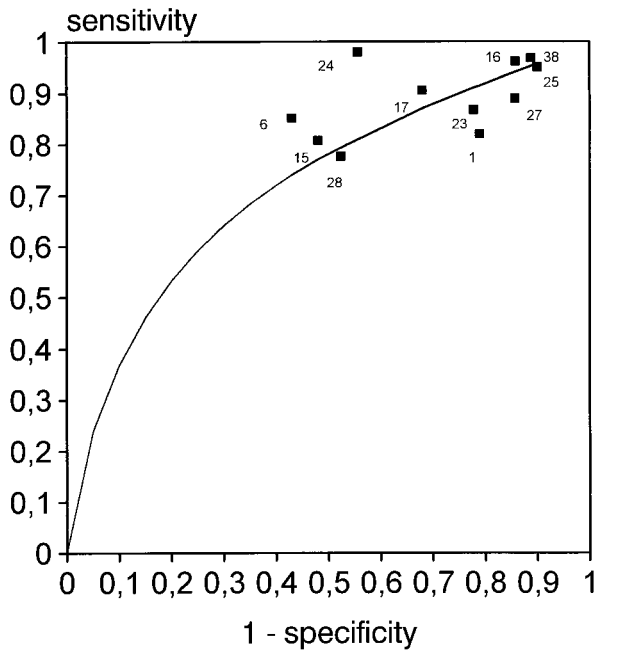

Figure 3. Summary ROC curve for SLR of the 11 pooled studies (thick line = fitted curve; thin line = extrapolated curve).

the external validity of the study in more detail. Agreement between both assessors was moderate, but poor for four items. These are quite important items because three of them were shown to indicate potential sources of bias.

Only 11 studies provided enough data to reconstruct the diagnostic table. If an important internal validity criteria, such as the presence of an accepted reference test, presence of independency of assessment, and absence of verification bias, should have been the minimum conditions for selection, no study would have entered the meta-analysis. All studies were prone to verification bias, as nearly all studies were retrospective studies based on data obtained in surgical patients. The assumption was made that patients who did not undergo surgery would have shown proportionally different positive and negative results for the SLR.[35] These patients could have been followed up for a definitive diagnosis at a later stage.[36] As already mentioned, the presence of verification bias makes a difference for the diagnostic accuracy of the SLR test.

All studies are surgical case-series taken from clinical practice in hospitals and are, as such, a mirror of that reality. The performance of the SLR test under ideal conditions in a prospective study based on a valid study protocol was not assessed. Moreover, the diagnostic accuracy of the test at other levels of medical care with a different patient-mix is not known. Although most basic demographic information is present in almost all studies, important information about duration of illness, comorbidity, and execution of the SLR is often lacking.

The two parameters for test accuracy, sensitivity and specificity, are highly heterogenous, but no clear outliers can be distinguished. One would hesitate to pool these studies based on these two parameters. Combining both into one, the diagnostic odds ratio revealed an outlier study. Given the obvious verification bias, this study could be excluded from the analysis, which resolved the problem of statistical heterogeneity of the ln(DOR).

The diagnostic odds ratio of the SLR was not dependent on the cutoff point used in the different studies. It decreased with a more valid design (blind assessment in both directions), a more homogenous casemix (by excluding patients who previously had had surgery), and year of publication. The outlier study with an obvious verification bias increases the DOR. All criteria outcomes referring to a more valid design demonstrate a lower DOR.

In half of the studies, the CSLR was evaluated in a selected group of patients without obvious selection criteria. The increased specificity results in a slightly higher positive predictive value, although the prevalence in this group of studies was slightly lower compared with the pooled SLR studies. If the CSLR would be used at this level of care as a confirmatory test in case of a positive SLR, the predictive value of a positive test would increase up to 0.94.

From the studies available for the statistical pooling, only four score positive in more than 50% on the total validity score. There is no strong association, however, between the total validity score and the diagnostic accuracy of the SLR.

The findings are limited by the poor quality of the primary studies and the poor reporting in the publications. Diagnostic studies should be planned on the basis of a research proposal, taking into account the basic

criteria for valid diagnostic research. As it is, information about the patient-mix was incomplete, information about the diagnostic path was not clear, and probably not all patients who have been tested were referred for surgery and thus were not included in most of these studies. Although the studies may be reflecting everyday clinical practice, they do not enable a valid evaluation of the diagnostic accuracy of the SLR.

The previous systematic reviews were not limited to the (C)SLR test, but also describe history and physical examination of patients with low back pain.[18,41] The authors concluded that physical examination, including the SLR, was only moderately accurate in diagnosing low back pain. The present analysis was confined to the SLR because the evidence on other assessments of neurologic compression by testing motor, reflex, and sensory functions is too limited to enable a quantitative analysis. The publications provided insufficient data to enable an evaluation of the added value of either foot-dorsiflexion or neck flexion.

One of the previous reviews suggested that the variance of the SLR could be explained partly by the variance of the cutoff point, whereas the present study found no association between the DOR and the cutoff point.[41] Sensitivity and specificity of the SLR were similar to the present study, although based on only five surgical studies. Another review on herniated lumbar discs also mentioned the problem of highly selected populations and its effect on the diagnostic accuracy and predictive values of the different tests.[3] Because of the nature the different tests providing traction on the nerve roots, the authors expected these tests to be sensitive but unspecific as to the cause of nerve irritation. The limited accuracy of the tests, SLR included, indicates the need for other confirmatory tests.

Two other nonsystematic reviews gave some figures for the accuracy of the SLR test based on a limited number of publications[2,11]: a moderate sensitivity of 0.80 and a specificity of 0.40, respectively, which is slightly different from the estimations of the present study. One of these reviews assumed the SLR to be most appropriate for testing the lower lumbar nerve roots.[11] In the present review, only few studies provided information about the ratio between lower and upper lumbar hernias, which limited the possibilities to address this question.

In summary, the results of this review demonstrate the limitations of the published studies in this field. Flaws in the design of these studies include independency of interpretation, verification bias and retrospective design. The highly selected populations at a high level of the medical care system limits the generalizibility of findings to primary care. Lack of information about the patients included in the study and the diagnostic process also limits the external validity of the results. Diagnostic research should evaluate the validity of the complete diagnostic process and study the evidence of the added value of the different tests used. Diagnostic research should be performed at all levels of clinical care to give an insight into the accuracy of a diagnostic test in different casemixes. The absence of primary care studies and evidence about the reproducibility of the tests confirms the urgent need for such studies.

■ Key Points

- The straight leg raising test is only evaluated in surgical case-series at the nonprimary care level.
- The more valid studies show a limited diagnostic accuracy of the SLR.
- Diagnostic research should evaluate the added value of different tests in a diagnostic process.

References

1. Albeck MJ. A critical assessment of clinical diagnosis of disc herniation in patients with monoradicular sciatica. Acta Neurochir (Wien) 1996;138:40–4.
2. Andersson GBJ, Deyo RA. Sensitivity, specificity, and predictive value: A general issue in screening for disease and in interpretation of diagnostic studies in spinal disorders. In: Frymoyer JW, ed. The Adult Spine: Principles and Practice. 2nd edition. New York: Raven Press Ltd, 1996.
3. Andersson GBJ, Deyo RA. History and physical examination in patients with herniated lumbar discs. Spine 1996;21:10S–18S.
4. Aronson HA, Dunsmore RH. Herniated upper lumbar discs. J Bone Joint Surg Am 1963;45-A:311–7.
5. Boden SD, Davis, DO, Dina TS, Patronas NJ, Wiesel SW. Abnormal magnetic resonance scans of the lumbar spine in asymptomatic subjects. J Bone Joint Surg Am 1990;72-A:403–408.
6. Charnley J. Orthopaedic signs in the diagnosis of disc protrusion. With special reference to the straight leg raising test. Lancet 1951;260:186–92.
7. Colditz GA, Burdick E, Mosteller F. Heterogeneity in meta-analysis of data from epidemiologic studies: A commentary. Am J Epidemiol 1995;142:371–82.
8. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.
9. Deyo RA, Tsui-Wu JY. Descriptive epidemiology of low back pain and its related medical care in the United States. Spine 1987;12:264–8.
10. Deyo RA, Loeser JD, Bigos SF. Herniated lumbar intervertebral disk. Ann Intern Med 1990;112:598–603.
11. Deyo RA, Rainville J, Kent DL. what can the history and physical examination tell us about low back pain? JAMA 1992;268:760–5.
12. Edgar MA, Park WM. Induced pain patterns on passive straight-leg raising in lower lumbar disc protrusion. J Bone Joint Surg Br 1974;56-B:658–67.
13. Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap test accuracy. Am J Epidemiol 1995;141:680–9.
14. Fleiss JL. Statistical methods for rates and proportions. New York: Wiley, 1973:143–7.
15. Gurdjian ES, Webster JE, Ostrowski AZ, Hardy WG, Lindner DW, Thomas LM. Herniated lumbar intervertebral discs: An analysis of 1176 operated cases. J Trauma 1961;1:158–76.
16. Hakelius A, Hindmarsh J. The significance of neurological signs and myelographic findings in the diagnosis of lumbar root compression. Acta Orthop Scand 1972;43:239–46.
17. Hirsch C, Nachemson A. The reliability of lumbar disk surgery. Clin Orthop 1963;29:189–95.
18. Hoogen JMM van den, Koes BW, Eijk JThM van, Bouter LM. On the accuracy of history, physical examination, and erythrocyte sedimentation rate in diagnosing low back pain in general practice. Spine 1995;20:318–27.
19. Hoogen JMM van den, Koes BW, Devillé W, Eijk JThM van, Bouter LM. The interobserver reproducibility of Lasègue's sign in patients with low back pain in general practice. Br J Gen Pract 1996;46:727–30.
20. Irwig L, Tosteson ANA, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120:667–76.
21. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic accuracy. J Clin Epidemiol 1995;48:119–30.
22. Irwig L, Glasziou P, Weijden T van der and the Cochrane Methods Working Group on Meta-analysis of diagnostic and screening tests. Intern document revised at the Third Cochrane Colloquium, Oslo 1995.
23. Jönsson B, Strömqvist B. The straight leg raising test and the severity of symptoms in lumbar disc herniation. A preoperative and postoperative evaluation. Spine 1995;20:27–30.

24. Kerr RSC, Cadoux-Hudson TA, Adams CBT. The value of accurate clinical assessment in the surgical management of the lumbar disc protrusion. Neurol Neurosurg Psychiatry 1988;51:169–73.
25. Knuttson B. Comparative value of electromyographic, myelographic and clinical–neurological examinations in diagnosis of lumbar root compression syndrome. Acta Orthop Scand 1961;49(Suppl.): 7–135.
26. Kortelainen P, Puranen J, Koivisto E, Lahde S. Symptoms and signs of sciatica and their relation to the localization of the lumbar disc herniation. Spine 1985;10:88–92.
27. Kosteljanetz M, Espersen JO, Halaburt H, Miletic T. Predictive value of clinical and surgical findings in patients with lumbago-sciatica. A prospective study (part I). Acta Neurochir 1984;73:67–76.
28. Kosteljanetz M, Bang F, Schmidt-Olsen S. The clinical significance of straight-leg raising (Lasègue's Sign) in the diagnosis of prolapsed lumbar disc: Interobserver variation and correlation with surgical finding. Spine 1988;13: 393–5.
29. Laird NM, Mosteller F. Some statistical methods for combining experimental results. Int J Technol Assess Health Care 1990;6:5–30.
30. Lau J. Meta-test version 0.6. Boston, MA: New England Medical Center, 1997.
31. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations. Stat Med 1993;12:1293–316.
32. Mulrow CD, Linn WD, Gaul MK. Assessing quality of a diagnostic test evaluation. J Gen Intern Med 1984;252:2418–22.
33. Nachemson A. The lumbar spine: An orthopedic challenge. Spine 1976;1: 59–71.
34. Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. Acad Radiol 1995;2:S48–56.
35. Schouw van der YT, Dijk van R, Verbeek ALM. Problems in selecting the adequate patient population from data files for assessment of new diagnostic tests. J Clin Epidemiol 1995;48:417–22.
36. Schouw van der YT, Verbeek ALM, Ruijs SHJ. Guidelines for the assessment of new diagnostic tests. Invest Radiol 1995;30:334–40.
37. Shiqing X, Quanzhi Z, Dehao F. Significance of the straight leg raising test in the diagnosis and clinical evaluation of lower lumbar intervertebral-disc protrusion. J Bone Joint Surg Am 1987;69-A:517–22.
38. Spangfort EV. The lumbar disc herniation. A computer-aided analysis of 2504 operations. Acta Orthop Scand 1972;142(Suppl.): 5–79.
39. Thompson SG, Pocock SJ. Can meta-analyses be trusted? Lancet 1991;338: 1127–30.
40. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ 1994;309:1351–5.
41. Vroomen PCAJ. The diagnosis and conservative treatment of sciatica. Chapter 2: A systematic review of the value of history and physical examination [dissertation]. Maastricht: Maastricht University, 1998:19–28.
42. Vucetic N, Svensson O. Physical signs in lumbar disc hernia. Clin Orthop 1996;333:192–201.
43. Waddell G, Allan DB, Newton M. Clinical evaluation of disability in back pain. In: Frymoyer JW, Ducker TB, eds. The Adult Spine. Principles and Practice. Eds. New York: Raven Press, 1991:164–5.
44. Weisel SE, Tsourmas N, Feffer H, Citrin CM, Patronas N. A study of computer-assisted tomography, I: The incidence of positive CAT scans in an asymptomatic group of patients. Spine 1984;9:549–51.
45. White AA, Gordon SL. Synopsis: Workshop on idiopathic low back pain. Spine 1982;7:141–9.

Address reprint requests to

W. Devillé, MD
EMGO Institute
Faculty of Medicine
Vrije Universiteit Amsterdam
van der Boechorststraat 7
1081 BT Amsterdam
The Netherlands
E-mail: w.deville.emgo@med.vu.nl